LITERATURE ON STATISTICAL DISTRIBUTIONS

(a proposal of an Information Retrieval System).

by Samuel Kotz
Temple University

Norman L. Johnson
University of North Carolina

Paper presented at the 37-th Session of the International Statistical Institute, London Sept. 3-11, 1969.

1. [illegible] for public release and sale; its distribution is unlimited.

12

As it was pointed out in the Presidential Address at this session, in the last 6-7 years the statistical community has witnessed a number of successful attempts to organize and classify the vast amount of statistical literature which has increased at an ever accelerating rate in the last twenty years.

The pioneering works of Dr. Frank Haight, Professor Maurice Kendall, Dr. William Buckland, Professors Patil, Lancaster, Wold, Walsh, Olkin, Sanga and several others should be particularly noted in the connection.

If I have missed certain noteworthy names from this list and I probably have, it is of course not intentially but due to ignorance or absent-mindedness.

It seems, however, to the authors of the present paper that we should move one step further and that a computerized information retrieval system for statistical techniques and methodology is both feasible and desireable at the present stage of development.

In addition to the printed version of our paper appearing on pages 303-306 of the Contributed Papers Volume 1 which I will assume that the audience is familiar with, I would like to report on an experiment utilizing computer techniques of what hopefully will become at least the first stage of an Information Retrieval System for statistical distributions and their application. First, I would like to give the background for this work.

Since 1963 Dr. N. L. Johnson and I have been engaged in compiling A Compendium of Statistical Distributions, a three volume project, two of which are at the present time in the final stages of proofreading and the first volume is due to appear this month.

During the course of preparation of the Compendium, we have collected over 2000 reprints and xerox copies of various papers from over 200 publications, some from obscure publications dealing with the subject. Most of these reprints are accompanied by abstracts taken from Mathematical Reviews and/or Referativnyi Zhurnal, Zentralblatt fur Matematik, Statistical Abstracts and others. On the basis of preliminary and partial investigations, it is estimated that the major papers on Statistical Distributions are scattered in over 335 journals. I have with me a list of these journals. It should be noted, however, that as it is seen from Table B on page 305, the 12 basic journals contain over 60% of paper and the remaining 240 less than 40%.

It became evident in the ocurse of our research that this type of endeavor requires permanent up-dating and revision in order to justify the great effort involved and to assure the usefulness of this work for numerous users. We were therefore contemplating the establishment of a permanent center to increase the operational value of the collection. As the first stage towards this aim we decided to code the information from each of the available papers according to a classification to be described in a moment.

This first stage took about six months, and was performed by qualified graduate students with our assistance.

In connection with the process of coding we have the following general comments. To determine the content of an article it was necessary to read many of them completely. This was especially true when beginning a folder of articles on a distribution not yet coded. In retrospect, the first few distribution took a very long time to code. When more than five distributions

were completed (and more than 500 articles coded) the task of coding was more easily accomplished. The average rate is estimated at 15 min./per article. It is readily admitted that someone very experienced in the field could have done the coding more quickly. However, this would be done at the expense of any interest in the mathematical content of the papers and would be educationally unrewarding. The question also arises as to the advisibility (even the possibility) of coding at a more rapid pace for long periods of time (7 or 8 hours a day). It is believed that this attitude of coding at a more relaxed pace is in keeping with the motivation for this computerized file. Namely, someone makes an accurate list of the content of a large number of articles so that many can have access to this information without an enormous investment of time on the part of many.

The information taken from 2000 papers is now being processed in coded form on IBM cards, and we are ready in principle to proceed with the operational activities, and to supply interested institutions and individuals with information on any distribution and/or specific characteristic of the distribution such as mathematical properties, estimation procedure, etc., details of which will now be given.

Before discussing the details, however, I would like to point out that as the collection grows, the manual classification of cards to supply the requested information is planned to be replaced by a computer program to be written by Dr. G. Koch of the Biostatistics Department of U.N.C.

The computerized filing scheme will be constructed according to principles studied in the dissertation of G. Koch entitled The Design of Combinatorial Information Retrieval Systems for Files with Multiple-valued Attributes, Univeristy of North Carolina, Mimeo Series No. 552. The chief advantage of such a system is that the retrieval time for various information requests will be almost independent up to a certain upper bound of the size of the file (i.e., the number of references to be included in the bibliography) which makes the updating rather a painless and routine task. This aspect of the research is considered to be both of an applied and basic nature. The basic aspect is related to the choice of the algebraic scheme from which the system is to be derived and then to the discovery of the most efficient way of implementing it in the computer. The applied aspect is that the resulting computerized system will be applied to a large and complex bibliography-namely that of statistical distributions.

However, even after the first stage we are already in possession of a rather unique and efficient classification procedure.

I will now give the details of our classification system:

The 80 columns of an IBM card are subdivided in the following manner:

Columns 1-3 — Journal identification number

Columns 7-10 — First page of paper

Columns 11-60 — assigned to distributions (see next page) are coded as follows:

0 if distribution is not discussed
1 if distribution is mentioned
2 if distribution is primary subject

Columns 61-78 — assigned to topics (see next page) are coded as follows:

0 if topic is not discussed
1 if topic is mentioned
2 if topic is primary subject

Column 79 — assigned to number of pages is coded as follows:

0 if 1-4 pages
1 if 5-8 pages
2 if 9-12 pages
3 if 13-16 pages
4 if 17-20 pages
5 if 21-24 pages
6 if 25-35 pages
7 if 36-50 pages
8 if more than 50 pages
9 if unknown

Column 80 — assigned to the language of the paper is coded as follows:

1 if English
2 if Russian
3 if French
4 if German
5 if Spanish
6 if Italian

The list of distribution families corresponding to columns 11-60:

11. Compendia and Bibliographical Sources
12. General Systems of Discrete Distributions
13. Binomial
14. Poisson
15. Geometric
16. Negative Binomial - (compound Poisson - Pascal)
17. Hypergeometric
18. Logarithmic Series
19. Compound and Generalized Discrete Distributions
20. Contagious Distributions
21. Miscellaneous Discrete
22. Multivariate Discrete Distributions
23. General Systems of Continuous Distributions
24. Normal (Gaussian)
25. Lognormal
26. Inverse Gaussian
27. Cauchy
28. X^2
29. Gamma
30. Exponential and Exponential type
31. Pareto
32. Weibull
33. Extreme Value - Gaumbel - Frechet's distributions
34. Logistic
35. Laplace - (double exponential)
36. Beta
37. Rectangular (uniform) and related distributions
38. F (and z)
39. t
40. Noncentral X^2
41. Quadratic Forms in Normal Variables
42. Noncentral F
43. Noncentral t
44. Generalized X^2 t and F (under non-standard normal assumptions)
45. Distributions of Correlation Coefficients
46. Miscellaneous Continuous Distributions
47. General Multivariate Distributions and Surfaces (Bivariate)
48. General Multivariate Distributions and Surfaces (Multivariate)
49. Multivariate normal (Bivariate)
50. Multivariate normal (Trivariate)
51. Multivariate normal (Multivariate)
52. Multivariate t-
53. Multivariate extreme-value
54. Multivariate exponential and Weibull
55. Multivariate Gamma

56. Wishart
57. Non-central Wishart and distribution of latent roots and vectors
58. Multivariate Beta and F
59. Non-central Multivariate Beta
60. Miscellaneous Multivariate Distributions

The list of topics corresponding to columns 61-78:

61. Origin and historical remarks
62. Definition, Distribution function, Characterizations
63. Moments, cumulants and other characteristics (excluding order statistics)
64. Genesis in models
65. Tables
66. Nomographs and Probability papers
67. Approximations to the distribution
68. Limiting forms
69. Transformation and relations to other distributions
70. Order statistics
71. Mathematical properties
72. Point estimation
73. Sequential estimation
74. Interval estimation
75. Test on parameters
76. Goodness of fit
78. Applications in statistical methodology
79. Application in sciences

This system is geared to reply to queries of the type: "list all the papers dealing with estimation methods of the shape parameter of the Weibull distribution." We would be able to supply up-to-date information on such a request without much redundancy. One of the major difficulties in our classification system, however, is that for its most efficient functioning it is desireable that the topics and distributions be mutually exclusive, otherwise we may over-supply with extraneous information.

A minor problem which has not been satisfactorally solved yet (besides the problem of possible missing information in the coded articles) is how to code the ten digit identification number for non-journal articles from various research centers and selections of books.

I would like to save the remaining time allocated by the Chairman for questions and especially for a discussion period. In particular I am very interested in your views about the practicality and usefulness of the proposed system in your statistical activities.

From private conversations with a number of distinguished delegates active in bibliographical statistical research as well as people who sponsor this work, I have discovered substantial interest in it, and while some of them expressed doubts whether the time is ripe for this rather sophisticated procedure of storing and utilizing information on statistical distributions and suggested that in their opinion the conventional method of books devoted to particular distributions or to particular topics of distribution theory is just as efficient and usable for the time being. Many others, however, were in full agreement that the delay from year to year increases the danger that in the not too distant future, when the information explosion reaches a certain saturation point, the inevitable task of the initialorganization and subsequent continuity of operational efficiency and smoothness of such a computerized system will be significantly more difficult and complex.

Acknowledgements

This research is supported in part by the United States Air Force Office of Scientific Research under the contract AF-AFOSR-69-1411 and in part by a grant from the same Office under the contract AF-AFOSR-68-1415.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Temple University
Department of Mathematics
Philadelphia, Pennsylvania 19122

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

LITERATURE ON STATISTICAL DISTRIBUTIONS. A PROPOSAL OF AN INFORMATION RETRIEVAL SYSTEM

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Scientific Interim

5. AUTHOR(S) *(First name, middle initial, last name)*
Samuel Kotz
Norman L. Johnson

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1970	10	

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
AFOSR 68-1411	
b. PROJECT NO. 9769-06	
c. 6144501F	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* AFOSR 70-1853 TR
d. 681304	

10. DISTRIBUTION STATEMENT

1. This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
TECH, OTHER	Air Force Office of Scientific Research (SRM) 1400 Wilson Boulevard Arlington, Virginia 22209

13. ABSTRACT

This paper presents an approach to classifying the literature on statistical distributions. The distribution families and other characteristics of the distributions and papers that were employed in the compilation of a compendium of statistical distributions are listed. Of over 2000 reprints and papers that were collected, 60 were published in 12 basic journals, the remainder being distributed over approximately 200 journals. Approximately 15 minutes were needed by a subject specialist to code each paper.

DD FORM 1 NOV 68 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information retrieval						
Statistical distributions						
Classification system						
Indexing						